BAND-ADAPTIVE FORMANT FREQUENCY ESTIMATION FROM NOISY SPEECH IN CORRELATION DOMAIN

Shaikh Anowarul Fattah, Wei-Ping Zhu, and M. Omair Ahmad
Dept. of Electrical and Computer Engineering
Concordia University, 1455 De Maisonneuve Blvd. W., Montreal, Quebec, Canada H3G 1M8

1. INTRODUCTION

Formant frequency is one of the most important speech parameters, which helps in better understanding human speech production mechanism. Estimation of formant frequencies from observed speech signal has applications in several areas, such as, speech synthesis, recognition, and compression. As an acoustic feature, formant offers phonetic reduction in speech recognition and it plays a vital role in the design of some hearing aids [1]–[3]. Formants are associated with peaks in the smoothed power spectrum of speech. Most of the formant estimation methods, so far reported, deal only with noise-free environments [1], [3]. For example, the formant estimators based on the linear predictive coding (LPC) or autocorrelation function (ACF) exhibit significant performance degradation in the presence of noise. Formant estimation from noisy speech is very difficult but essential for practical applications. In [2] and [4], formant frequency estimation methods have been proposed in order to handle noisy environments. The method in [2] utilizes an adaptive filter bank (AFB) and its performance depends on initial estimates. The method in [4] introduces a residual optimization technique based on a correlation model of voiced speech signals.

The objective of this paper is to estimate formant frequencies accurately under a severe noisy condition. In order to overcome the detrimental effect of noise, instead of using the conventional ACF, we introduce a band-limited repeated ACF (RACF) of the observed noisy speech. It has been shown that the RACF is pole-preserving and capable of drastically reducing the effect of noise. First, a band-adaptive filter-bank is employed on a zero lag compensated ACF to separate each formant frequency region prior to the formant estimation. Autocorrelation operation is then repeated on each of the resulting band-limited ACFs. Finally, a spectral peak picking method is employed to each of the band-limited RACFs to extract formant frequencies. The proposed algorithm is tested on synthetic and natural speech signals in the presence of noise and experimental results demonstrate a very satisfactory performance.

2. PROPOSED METHOD

2.1 Formant Estimation in Correlation Domain

The overall human vocal-tract can be represented by a P-th order all-pole system with a transfer function given by

$$H(z)=\frac{G}{A(z)}=\frac{G}{1-\sum_{i=1}^{P}a_i z^{-i}}=\frac{G}{\prod_{i=1}^{P}(1-p_i z^{-1})} \quad (1)$$

where G is a gain factor, $\{a_i\}$ the autoregressive (AR) system parameters, and p_k the system pole. Free resonances of the vocal tract system are called formants. In order to model each formant, a pair of complex conjugate poles is required. Formant frequency (F_k) and bandwidth (B_k) can be computed from the pole magnitude r_k, angle ω_k, and sampling frequency F_S as

$$F_k=\omega_k(F_S/2\pi)\;;\qquad B_k=-(F_S/\pi)\ln(r_k) \quad (2)$$

Filtering speech signal $x(n)$ by the vocal-tract filter $A(z)$, one can obtain an error or residual signal and minimizing the mean error leads to following relation

$$r_x(\tau)=\sum_{k=1}^{P}a_k r_x(\tau-k),\ \tau=1,2,\ldots\ldots,L \quad (3)$$

where an estimate of the ACF $r_x(\tau)$ of N length data $x(n)$ can be computed as

$$r_x(\tau)=\frac{1}{N}\sum_{n=0}^{N-1-|\tau|}x(n)x(n+|\tau|),\quad \tau=0,1,\ldots\ldots,L'-1;L'\le N \quad (4)$$

It is to be noted that the ACF $r_x(\tau)$ is a pole preserving function as seen from its z-transform for nonnegative lags

$$R_x(z)=\sum_{k=1}^{P}\frac{D_{1k}}{1-p_k z^{-1}} \quad (5)$$

where $\{D_{1k}\}$ are partial fraction coefficients. Estimating parameters $\{a_i\}$ from the least-squares (LS) solution of (3), system poles as well as formants are extracted. In order to reduce the estimation variance, a combination of more than P equations can be used in the LS estimation.

2.2 Repeated Autocorrelation in Noise

Under noisy condition, observed speech is given by

$$y(n)=x(n)+v(n) \quad (6)$$

where, the additive noise $v(n)$ is assumed to be zero mean with variance σ_v^2. The ACF of $y(n)$ can be expressed as

$$\begin{aligned} r_y(\tau)&=r_x(\tau)+r_w(\tau)\\ r_w(\tau)&=r_v(\tau)+r_{vx}(\tau)+r_{xv}(\tau)\end{aligned} \quad (7)$$

Here the ACF $r_v(\tau)$ of $v(n)$ mainly affects only the zero lag and the effect of crosscorrelation terms are negligible. Thus, the effect of noise in the correlation domain is relatively less pronounced than that in the signal domain. If the autocorrelation operation is repeated on $r_y(\tau)$, in the resulting once-repeated ACF (RACF), the effect of noise term will be drastically reduced. It can be shown that the RACF, like the ACF, preserves the poles of the vocal-tract AR system.

Since the effect of $r_v(\tau)$ is mainly pronounced at the zero lag, in the computation of the RACF, $r_y(0)$ is excluded, which also offers a significant noise reduction. Instead of conventional ACF, the RACF can be used in (3) to estimate the system poles as well as formants. However, in this case, under severe noisy condition, there is a tendency of missing weak poles in the LS solution. In order to avoid such a situation, in what follows we propose a scheme where formant estimation is performed in a band-limited region utilizing the repeated autocorrelation.

2.3 Band-pass Filtering

In order to extract each formant frequency accurately under noisy condition, instead of using directly the RACF, we propose to employ a two stage technique. First, an adaptive filter-bank is employed on the zero lag compensated ACF of the observed speech to separate each formant frequency region in the correlation domain prior to the formant estimation. The autocorrelation operation is then repeated on the resulting band-limited filtered ACF in order to achieve the advantageous features of the RACF. Thus, instead of estimating all formants together from the RACF described in Sec. 2.2, we propose to estimate each formant separately from a band-limited RACF. One important advantage of using filter-bank is that in each of the band-limited RACFs the energy of the neighboring formants is greatly reduced and it contains energy primarily from only one formant. The band-pass filters also attenuate the energy at the pitch frequency in their outputs. Thus, use of band-limited RACFs is expected to offer a better noise immunity at low levels of signal to noise ration (SNR). In order to estimate three formants, in the filter-bank, three band-pass filters are designed depending on the conventional frequency range of the formant. These filter coefficients are then used to filter the compensated ACF into three band-limited spectral regions. The band-pass filter frequency regions are updated over time based on the previous formant frequency estimates in order to obtain a smooth formant track. The first three formant frequencies of voiced speech segments are estimated from the three band-limited RACFs using spectral peak picking method. In this case, the search zone for estimating a formant frequency is restricted within the band-pass filter frequency region. Finally, from each band-limited RACF, one formant frequency is estimated.

3. SIMULATION RESULTS

The proposed formant frequency estimation algorithm has been tested using various synthetic vowels synthesized using the Klatt synthesizer [1] and some natural vowels extracted from the North-Texas standard databases [1], [5]. For the performance comparison, the 12th order LPC [1] and the AFB methods [2] are considered and the percentage root-mean-square error (RMSE) at different noise levels are computed where each noise level consists of 20 independent trials of noisy environments. In our implementation, we perform the formant estimation every 10 ms with a 20 ms window applied to overlapping voiced speech segments. In literature, the region of formant frequencies has been well-studied [1]. In the filter-bank, for the first formant fourth order and for other formants sixth order butterworth filters have been used [1]. In Table 1, the estimated %RMSE(Hz) is shown for two synthesized vowels at SNRs 0 dB and 5 dB. It is found that the proposed method provides lower %RMSE (Hz) for both male and female speakers. In Table 2, %RMSE (Hz) for natural vowels /a/ and /i/ (contained in words "hod" and "heed") is shown, which indicates a better estimation accuracy obtained by the proposed method.

Table 1. %RMSE (Hz) for Synthetic Vowels

Vowels			0 dB			5 dB		
			Prop.	LPC	AFB	Prop.	LPC	AFB
Male	/a/	F1	10.49	23.63	31.29	7.81	15.87	11.74
		F2	10.36	27.78	34.82	3.92	15.53	9.51
		F3	10.34	19.28	19.34	6.13	13.19	8.23
	/i/	F1	21.42	28.53	28.16	13.76	19.28	16.25
		F2	4.01	9.68	4.27	2.71	7.54	3.33
		F3	5.25	13.29	7.75	3.68	7.82	3.97
Female	/a/	F1	11.43	17.76	16.76	5.33	9.76	15.67
		F2	9.22	19.43	14.39	5.12	8.68	7.49
		F3	3.71	9.26	4.57	2.09	3.18	2.34
	/i/	F1	23.31	39.81	32.27	15.48	28.27	19.14
		F2	10.93	26.21	13.78	5.27	12.43	6.19
		F3	3.12	15.83	3.78	2.32	7.63	2.78

Table 2. %RMSE (Hz) for Natural Vowels

Vowels			0 dB			5 dB		
			Prop.	LPC	AFB	Prop.	LPC	AFB
Male	/a/	F1	10.27	14.33	13.93	7.23	9.57	8.54
		F2	16.76	44.93	28.76	13.12	28.27	16.29
		F3	14.24	38.01	23.34	11.43	23.67	18.31
Female	/i/	F1	10.17	21.19	16.84	5.47	8.61	5.95
		F2	11.28	23.78	19.49	6.75	11.28	10.21
		F3	13.31	31.29	24.82	5.22	21.92	14.58

4. CONCLUSION

The proposed scheme offers an attractive feature that it estimates formant frequencies from band-limited formant regions utilizing advantages of the RACF. It has been clearly observed from experimental results on synthetic and natural speech signals under noisy conditions that the proposed method provides a high degree of estimation accuracy at a moderate to low levels of SNR.

REFERENCES

[1] D. O'Shaughnessy, (2000). Speech Communications: Human and Machine (2nd ed.). *IEEE Press, NY*.

[2] K. Mustafa and I. C. Bruce, (2006). Robust formant tracking for continuous speech with speaker variability. IEEE Trans. Audio Speech Lang. Processing, 14, 435–444.

[3] L. Deng, A. Acero, and I. Bazzi, (2006). Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint. IEEE Trans. Audio Speech Lang. Processing, vol. 14, no. 2, pp. 425–434, Mar. 2006.

[4] S. A. Fattah, W.-P. Zhu, and M. O. Ahmad, (2007). An approach to formant frequency estimation at low signal-to-noise ratio. ICASSP'07, 4, 469–472.

[5] J. M. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.*, 97, 3099–3111.